Toward automatic improvement of language produced by non-native language learners

Creutz, Mathias

Linköping University Electronic Press
2019-09-30

Creutz , M & Sjöblom , E I 2019 , Toward automatic improvement of language produced by non-native language learners . in D Alfter , E Volodina , L Borin , I Pilán & H Lange (eds) , Proceedings of the 8th Workshop on Natural Language Processingfor Computer Assisted Language Learning (NLP4CALL 2019) . Linköping Electronic Conference Proceedings , no. 164 , NEALT Proceedings Series , no. 39 , Linköping University Electronic Press , Linköping , pp. 20-30 , Workshop on Natural Language Processing for Computer Assisted Language Learning , Turku , Finland , 30/09/2019 .

http://hdl.handle.net/10138/306947

Toward automatic improvement of language produced by non-native language learners

Mathias Creutz **Eetu Sjöblom**
Department of Digital Humanities
Faculty of Arts
University of Helsinki
Unioninkatu 40, FI-00014 University of Helsinki, Finland
{mathias.creutz, eetu.sjoblom}@helsinki.fi

Abstract

It is important for language learners to practice speaking and writing in realistic scenarios. The learners also need feedback on how to express themselves better in the new language. In this paper, we perform automatic paraphrase generation on language-learner texts. Our goal is to devise tools that can help language learners write more correct and natural sounding sentences. We use a pivoting method with a character-based neural machine translation system trained on subtitle data to paraphrase and improve learner texts that contain grammatical errors and other types of noise. We perform experiments in three languages: Finnish, Swedish and English. We experiment with monolingual data as well as error-augmented monolingual and bilingual data in addition to parallel subtitle data during training. Our results show that our baseline model trained only on parallel bilingual data sets is surprisingly robust to different types of noise in the source sentence, but introducing artificial errors can improve performance. In addition to error correction, the results show promise for using the models to improve fluency and make language-learner texts more idiomatic.

1 Introduction

It is difficult to express oneself well in a new language. Language students can learn grammar and vocabulary by filling in blanks in carefully prepared exercise sentences, but the students also need to practice speaking and writing in realistic scenarios. When students write their own texts, they need corrective feedback. We are interested in finding out to what extent computers can provide the necessary corrections, a task traditionally performed by human teachers. However, human teachers are not always available and the students will want to carry on using the language outside the language class. A tool helping language learners to produce more correct and more natural sounding expressions can enhance the learning process and encourage the students to use the new language in real situations. In addition, findings since the 1980s suggest that language students that receive corrective feedback from computers rather than human teachers learn better and perceive the feedback as more neutral and encouraging (Behjat, 2011).

In this paper, we study automatic paraphrasing methods on sentences produced by learners of three languages: Finnish, Swedish and English. A paraphrase is an alternate way of expressing a meaning using other words than in the original utterance, such as the sentence pair: *"Why don't you watch your mouth?"* ↔ *"Take care what you say."*

Our goal is to discover to what extent we can improve the spelling, grammar and naturalness of text written by non-native language users. We are not primarily interested in creating spell or grammar checkers, but we are interested in seeing whether it is possible to make "noisy" non-standard sentences sound more natural. Non-native users may be struggling to find fluent, natural sounding idiomatic expressions. Paraphrase generation may be a way to "translate" sentences produced by language learners to sentences that are grammatically correct and sound more authentic to native speakers.

In the present work, we do not set out to explicitly mark the errors made by the learners or suggest corrections to each of the errors separately. Rather, for each sentence produced by the non-

native language user, we propose an alternative, corrected sentence. The proposed sentence can differ significantly, or not at all, from the original sentence, depending on the quality of the original input. By comparing the original and altered sentence, the language learner can identify errors and learn new expressions.

Our work is closely related to the field of grammatical error correction (GEC), although our focus is broader. We are not only interested in *grammar*, but also in fluency and naturalness in a broader sense. Furthermore, the concepts of *error* and *correction* are too narrow, in our opinion, since we are interested in better, or more effective, ways of conveying a message.

Nonetheless, from our point of view, GEC can provide us with useful data sets, methods, as well as evaluation guidelines and metrics. Dahlmeier et al. (2013) introduce the NUCLE corpus, which was used in the CoNLL-2014 shared task on Grammatical Error Correction (Ng et al., 2014). NUCLE is an annotated corpus of English texts written by non-native English speakers. Twenty-eight error types have been annotated manually, such as incorrect preposition or verb tense. Napoles et al. (2017) present JFLEG, an English parallel corpus incorporating fluency edits, in order not only to correct grammatical errors but also make the original text more native sounding. Anastasopoulos et al. (2019) add Spanish translations to the JFLEG corpus.

Grammatical error correction systems are typically evaluated using metrics that compare the corrections suggested by the system to a set of gold standard corrections. The MaxMatch (M^2) algorithm (Dahlmeier and Ng, 2012) matches the system output to the gold standard and computes the sequence of edit operations that has maximal overlap with the gold standard annotation. This set of corrections is then scored using the F_1 measure. In the CoNLL-2014 shared task (Ng et al., 2014), the M^2 scorer is revised. In order to emphasize the precision of the suggested corrections twice as much as recall, the $F_{0.5}$ measure is used instead of F_1. Felice and Briscoe (2015) propose another metric, the I measure, which addresses some shortcomings of M^2, such as not distinguising between not proposing an edit versus proposing the wrong edit. Napoles et al. (2015) develop the Generalized Language Evaluation Understanding metric (GLEU) inspired by BLEU (Papineni et al., 2002), which seems to correlate better with the human ranking than the F and I measures.

When it comes to methods utilized in GEC, a broad range of approaches exist. The participants in the CoNLL-2014 shared task (Ng et al., 2014) propose systems based on classifiers (Naïve Bayes, averaged perceptron, maximum entropy), statistical language models, phrase-based and factored translation models, rule-based approaches, as well as combinations of these methods. More recently, machine translation has been the predominant framework. Sentences containing errors are translated into corrected sentences. Neural machine translation (NMT) generally requires large amounts of training data and has been shown to be sensitive to noisy data (Belinkov and Bisk, 2018). Therefore approaches have been suggested where "noise" of the desired characteristics are incorporated in the training data, such that the system learns to remove the noise in the translation (Belinkov and Bisk, 2018; Michel and Neubig, 2018; Anastasopoulos et al., 2019). Combining neural machine translation with statistical machine translation (SMT) is also claimed to produce better results (Grundkiewicz and Junczys-Dowmunt, 2018). Furthermore, GEC can be studied as a low-resource machine translation task, where in addition to adding source-side noise other techniques are used: domain adaptation, a GEC-specific training-objective, transfer learning with monolingual data, and ensembling of independently trained GEC models and language models (Junczys-Dowmunt et al., 2018), noisy channel models (Flachs et al., 2019) and unsupervised SMT (Katsumata and Komachi, 2019).

We are interested in the Nordic languages Finnish and Swedish. In addition, we perform experiments on English data. We use neural machine translation to produce paraphrases of original sentences written by non-native language learners. We are especially interested in the low-resource scenario, where in-domain, task-specific training data is scarce or non-existent, which is the case with Finnish and Swedish. Our approach uses multilingual character-level NMT in combination with out-of-domain machine translation data to deal with the lack of task-specific data. The data sets used for training and testing are described in Section 2. Our machine translation model and training process are described in Section 3. We then turn to our experiments in Section 4. The

models are evaluated using qualitative analysis and manual annotation, and the results are described in Section 5. Finally we conclude with a discussion in Section 6.

2 Data

We test our models on genuine text produced by non-native language learners. For training we use a large collection of subtitles.

2.1 Test data

As our test data we use parts of the YKI Corpus.[1] The corpus has been compiled from the examinations of the Finnish National Certificates of Language Proficiency, which is a language testing system for adults. Examinations can be taken in nine languages: English, Finnish, German, Italian, North Sami, Russian, Spanish, and Swedish. There are three test levels (basic, intermediate and advanced), which offer six levels of proficiency (1–2, 3–4, 5–6). The corpus contains data from all nine languages and levels. The YKI corpus is intended for research purposes. Access is provided by request.[2]

For each of our languages of study (Finnish, Swedish, and English) we have extracted the texts produced by twelve different language learners at random.[3] We have used the so-called "new material (2011–)". The learners are on proficiency levels 1–2 and the writing assignments given to them are on the basic level. The texts in the data represent the genres "informal letter or message", "formal letter or message", "opinion", "feedback" and "announcement". Examples of three texts in the data are shown in Table 1. The full extracted Finnish set contains 376 unique sentences, Swedish 332, and English 315. The data sets do not contain corrected versions of the sentences.

The backgrounds of the learners of Finnish and Swedish are quite diverse, whereas the English data is produced by a more homogeneous group of people. The Finnish learners consist of nine women and three men. Their native languages are: Russian (3), English (2), Chinese (2), German (1), Spanish (1), Turkish (1), and other (2). Among the Swedish learners there are ten women and two men. Their native languages are: Finnish (3), English (2), Estonian (2), Russian (1), French (1), German (1), Thai (1), and other (1). The English learners consist of eight men and four women. Eleven are native Finnish speakers and one is a Swedish speaker.

2.2 Training data

Our models are trained on data extracted from subtitles from movies and TV episodes. Large numbers of subtitles have been collected from `http://www.opensubtitles.org/` and aligned across languages to produce the OpenSubtitles corpus (Lison and Tiedemann, 2016; Lison et al., 2018). We have used the parallel subcorpora English–Finnish (23 million sentence pairs), English–Swedish (15 million sentence pairs), and Finnish–Swedish (12 million sentence pairs). These corpora allow us to train multilingual machine translation systems between the three languages, but it is also possible to perform so-called "zero-shot" translation from one language to itself.

The style of the subtitle data is not a perfect match for our test data. However, the conversational nature of subtitles make them suitable for modeling dialogues and everyday colloquial language (Lison et al., 2018). Our test data is produced by language learners at a basic level, who are mostly trying to express themselves in everyday language. In that sense it makes sense to use OpenSubtitles as training data. Furthermore, the subtitles are not restricted to a narrow genre or domain. The movies and TV series span from lighthearted productions for toddlers to historic dramas targeting older audiences, involving quite varied and distinct vocabulary (Paetzold and Specia, 2016).

In some of our experiments we use additional, monolingual data from the Opusparcus corpus (Creutz, 2018). Opusparcus consists of sets of sentential paraphrases, that is, pairs of sentences in the same language that mean essentially the same thing. The paraphrases of Opusparcus have been extracted from the OpenSubtitles corpus, so this monolingual data is similar in style to our bilingual training data. We use the Finnish, Swedish and English subsets of Opusparcus.

[1] `http://yki-korpus.jyu.fi/?lang=en`

[2] E-mail: `yki-info@jyu.fi`

[3] The participant IDs are: fi: 75798, 75946, 76023, 76030, 76354, 76357, 76361, 76362, 76365, 80504, 85081, 86465; sv: 70094, 70096, 70489, 72570, 72919, 76606, 76686, 76757, 76758, 76759, 77974, 77975; en: 68079, 68112, 69336, 69632, 69635, 69874, 72098, 72099, 72262, 76537, 76705, 77616.

Moi Maija: Minä olen kiinassa lomamatkalla. Olen ollut Kiinassa kahden viikoon. Minä jo kävi monissa kaupungissa. Se oli tosi mukavaa matkaa. Tavataan paljon ystäviä. Olen syönyt paljon kiinalaista herkuiset ystäviäni mukaan. Tosi hauskaa! Minä vielä haluan käymään Shanghaissa ja ostan jotaikin Shanghaista. Toivottavasti, nähdään pian! Terveisin, Matti
Hejsan Tove! Nu har jag äntligen kommit till mitt nya stad Vaasa. Jag mår verkligen bra men litet trött är jag. Flyttningen till bostaden tog fyra timmar och de var två män som hjälpte mig att bära tunna möbler. Bostaden är ljus och här finns stora fönster som ger dagljus till rummet. Det finns två rum, kök och WC, 56m alltså ganska stor lokalen åt mig. Kom och hälsa mig nästa månad. Vi ska ringa. Varma hälsningar åt er alla, Maija
Dear Bob! Thank you for a gift. It was beatufull! You still remember even we haven't met for long time. We celebreat with family our home. Parents, brothers, sisters were there. Family things... We had one thing which I don't Forget never. We take a photo where were Mum and Dad, both sisters and my brother all together the one picture! All peoples same place. Awsome. Please visit to us Bob. I would like to see You very soon! Yours, Matti

Table 1: Examples of three texts of the genre "informal letter" from the YKI Corpus (fi, sv, en). All of these particular three texts contain errors, but in comparison the Swedish text seems to be on the most advanced level, followed by English and Finnish. Despite the errors the texts are intelligible.

3 Model and Training

We adopt the neural machine translation (NMT) approach to paraphrase generation, using a standard encoder-decoder architecture. In an encoder-decoder model, the encoder maps an input sentence to a sequence of continuous vectors. The decoder then generates an output sentence based on the vector representations. Multiple different encoder and decoder choices can be used in the overall encoder-decoder architecture. Architectures based on recurrent neural networks (Luong et al., 2015) or self-attention (Vaswani et al., 2017) are the most common.

For our experiments, we choose the Transformer model by Vaswani et al. (2017). The Transformer has achieved state-of-the-art results in NMT and has found wide use in different sequence-to-sequence problems. It is based solely on self-attention within the encoder and the decoder, as well as attention between the encoder and the decoder, discarding the recurrent connections found in many earlier NMT architectures (Bahdanau et al., 2014; Luong et al., 2015). We train all our models as character-based models in an attempt to make the models more robust to typos and other noise present in the data. For training the multilingual models, we follow Johnson et al. (2017) by prefixing each source sentence with a target-language flag.

The hyperparameter choices for the Transformer model follow the recommended setup of OpenNMT-py (Klein et al., 2017), which we use for all experiments. We use 6 layers in both the encoder and the decoder, hidden states and character embeddings with 512 dimensions with separate embeddings for the encoder and decoder, 8 attention heads, and a feed-forward network with 2048 dimensions within the layers. We use a dropout probability of 0.1 between layers. All models are trained for 300k steps or until convergence, with a validation score as the convergence criterion. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0001 and a token batch size of 4096. At inference time we use beam search with beam size 12 to produce the outputs.

4 Experiments

We perform experiments on translation models trained in five different setups. All setups are built on our baseline model, which can translate from any of the three languages Finnish, Swedish or English to any of the same three languages.

4.1 Baseline model trained on bitexts

Our baseline model is trained on all of the OpenSubtitles parallel data (bitexts) for the three languages. This amounts to a total of approximately 50 million unique sentence pairs. We use both directions for all language pairs, but do not train on monolingual data (that is, the source and target sentences are never in the same language). We use

this model to produce paraphrases in two ways:

i) Zero-shot translation within the same language: For instance, the model translates from Finnish to Finnish although it has never seen training data where both the source and target sentence have been Finnish sentences. However, the training data does contain Finnish source and target sentences, but always aligned with a sentence in another language.

ii) Pivoting via a second language: The source sentence is translated into another language and then back to the source language. For example, a Finnish sentence is translated into English and then back to Finnish.

4.2 Baseline + Clones

As the baseline model does not see monolingual data during training, paraphrases have to be generated either using zero-shot translation or pivoting. Because zero-shot translation generally suffers from lower performance compared to language pairs seen during training, we attempt to improve the model by adding monolingual data. We do this by simply using copies of sentences from the OpenSubtitles training sets in addition to the full parallel data. We randomly sample 10 million sentences per language and use the same sentence as the source and target during training.

4.3 Baseline + Opusparcus

Because we are interested in generating fluent and natural paraphrases for the input sentences, we also experiment using paraphrase pairs as monolingual data instead of cloned sentences. In this case the model sees alternative ways of formulating sentences, phrases and lexical items. An example of an English source/target pair is: *"He believes in you."* ↔ *"He has faith in you."* Our paraphrase pairs come from the Opusparcus paraphrase corpus. We use 20 million pairs for English, 3.5 million for Finnish, and 1.8 million for Swedish. These data set sizes have been shown to perform well in a paraphrase detection task in earlier work (Sjöblom et al., 2018).

4.4 Baseline + Error-augmented monolingual data

The OpenSubtitles data consists of mostly clean sentences and proper language, although some noise, such as misspellings or optical character recognition errors, is present (Tiedemann, 2016). This is in contrast to our test data, where the majority of sentences contain errors. In our fourth setup we introduce artificial noise to our training data in an attempt to improve performance on noisy test sentences. We sample one million sentences for each language from the OpenSubtitles data, and for each sentence generate an erroneous pair using two types of errors: 1) Typos are introduced by randomly deleting a character from a word, swapping two adjacent characters, inserting an extra character or duplicating a character. 2) Inflection errors are introduced by randomly changing the inflection of a noun or a verb within the sentence using the UralicNLP toolkit for Finnish (Hämäläinen, 2019) and HFST tools for English and Swedish (Lindén et al., 2013). We randomly introduce 1–3 errors from either category to each sentence. The erroneous sentence is used as the source and the original as the target during training. Examples 1 and 2 show source sentences with typos and inflection errors respectively, with the corresponding correct targets:

1. ***Ae*** *taskuussa näköjään* ***voittsa tikarrin*** *saappaassa.* → *Ase taskussa näköjään voittaa tikarin saappaassa.*

2. *After she* ***attacks*** *you, perhaps you* ***had see*** *her?* → *After she attacked you, perhaps you have seen her?*

4.5 Baseline + Error-augmented bilingual data

Finally, in an attempt to improve the pivot-based method without monolingual data, we augment bilingual data for all language pairs with errors. We sample one million sentences pairs for each language pair, and use the same sentence pairs for both translation directions. The pipeline for generating the erroneous data is identical to the previous setup. The source sentences contain artificially introduced errors, whereas the target sentences are correct, as in: *"I had to* ***got*** *the bigger* ***one's****."* → *"Piti saada isompi."*

5 Evaluation

Our test sets do not contain gold standard reference sentences, and therefore we cannot use automated metrics to evaluate our models. Instead we will attempt to analyze the output of our models qualitatively and we also perform manual annotation of the generated sentences in two of the setups.

5.1 Qualitative evaluation

As expected, the baseline model (Section 4.1) performs poorly in a zero-shot translation scenario. The model is generally unable to produce a paraphrase with the same semantic content as the source sentence, and many of the produced sentences contain artifacts that can be traced back to one of the other languages, and the multilingual nature of the model. Examples of such artifacts are producing mixed language or incorrectly translating false friends, such as: *"Siinä on Teidän perheen valokuva."* → *"Siinä on erään* ***familjen*** *valokuva."*, *"Thank you for a gift."* → *"Thank you for a* ***poison****."* (The Swedish word for *family* has been inserted into a Finnish sentence, and the English word *gift* means poison in Swedish.)

Pivoting through another language works better as the model now only needs to translate between language pairs explicitly trained on. Examples of the intermediate steps (pivot languages) and the final paraphrases can be seen in Table 2. Many of the errors in the original source sentences have been corrected, although some sentences retain incorrect sentence structure or word forms from the source. Distortion of the source sentence semantics can also be seen in some cases. In the pivot scenario we also deal with the problem of compounding errors because of the two separate translation steps.

We now turn to the models trained on monolingual data in addition to bilingual parallel data. A general trend emerges with all three models where monolingual data was used (Sections 4.2, 4.3, and 4.4). The models will most of the time simply copy the source, including the errors present in the sentence. While this is somewhat expected of the model where clones were used, it is surprising that even the model with paraphrase data exhibits this behavior. The Opusparcus paraphrase corpus does not contain pairs with identical source and target sentences. The error-augmented monolingual data seems to aid in correcting some typographical errors in the source sentences but does not correct bad inflection to the same extent. The sentence structure of the generated paraphrase is generally identical to that of the source sentence: *"I* ***wiss*** *that you move the other* ***plase*** *and you can sleep very well"* → *"I wish that you move the other place and you can sleep very well"*

Guided by the results from pivot-based methods and the attempts to use monolingual data in training, our final setup incorporates error-augmented bilingual data instead of monolingual data (Section 4.5). A look at the generated phrases does not reveal consistent improvements over the baseline model, as shown in Table 3. The baseline model already corrects most typos, and while there are examples of phrases where the baseline model generates an incorrect word or inflection and the error-augmented model a correct one, the converse is true in other cases. We will compare the quality of the two models using manual annotations in the next section.

5.2 Manual annotation

Based on the qualitative assessment in the previous section, we have chosen to manually annotate paraphrases generated by two models using the pivot-based method. The models selected for annotation are the baseline model as well as the model with added error-augmented bilingual data. Annotators were shown one sentence pair at a time. The annotation task was to compare the original sentence to the generated paraphrase and assess the correctness and semantic adequacy of the parapahrase using a single four-grade scale. The annotation categories were the following: 1 (Bad paraphrase, erroneous language), 2 (Mostly bad paraphrase, multiple errors), 3 (Mostly good paraphrase, minor errors), and 4 (Good paraphrase, correct language).

For English and Finnish, two independent annotations were collected for each paraphrase. The inter-annotator agreement as measured by Cohen's Kappa is 0.43 (Moderate) for English and 0.50 (Moderate) for Finnish. Only one person annotated Swedish and consequently no inter-annotator agreement score can be calculated.

The manual annotation results are shown in Table 4. The results show an overall trend of the error augmented model performing better. For all languages the percentage of phrases annotated as 1 decreases, that is, the models generate less completely incorrect paraphrases. On the other end of the scale, the percentage of phrases annotated as category 4 decreases slightly for English, increases very slightly for Finnish, and increases significantly for Swedish.

6 Discussion and Conclusion

We have shown that a straight-forward character-based neural machine translation model trained on

Hey, Mary. I'm in China on vacation. I've been in China in two weeks. I already went to many towns. It was a really nice trip. Meet a lot of friends. I've eaten a lot of Chinese friends with my delicious friends. That's really funny! I still want to go to Shanghai and buy something from Shanghai. I hope I'll see you soon! Hello, Matt. → *Hei, Mary. Olen Kiinassa lomalla. Olen ollut Kiinassa kahden viikon päästä. Kävin jo monissa kaupungeissa. Se oli mukava matka. Tässä on paljon ystäviä. Olen syönyt paljon kiinalaisia ystäviäni. Todella hauskaa! Haluan yhä ostaa jotain Shanghailta. Toivottavasti näen sinut pian! Hei, Matt.*
Hi, Tove! Now I've finally come to my new city Vaasa. I'm really fine, but I'm a little tired. The moving to the house took four hours and they were two men who helped me carry thin furniture. The house is light and here are big windows that give daylight to the room. There are two rooms, kitchen and kitchen, 56 metres of the local for me. Come and tell me next month. We're gonna call. Warm greetings for all of you. Maija. → *Hej, Tove! Nu har jag äntligen kommit till min nya stad Vaasa. Jag mår bra, men jag är lite trött. Att flytta till huset tog fyra timmar och de var två män som hjälpte mig att bära tunna möbler. Huset är ljust och här är stora fönster som ger dagsljus till rummet. Det finns två rum, kök och kök , 56 meter för mig. Kom och berätta nästa månad. Vi ringer. Varma hälsningar för er alla. Maja.*
Hyvä Bob! Kiitos lahjasta. Se oli hienoa! Muistat vieläkin, ettemme ole tavanneet pitkään. Juhlimme perhettämme. Vanhemmat, veljet, siskot olivat siellä. Perheasioita... Meillä oli yksi asia, jota en koskaan unohda. Otamme kuvan, missä äiti ja isä olivat, molemmat siskoni ja veljeni yhdessä. Kaikki ihmiset samaan paikkaan. Mahtavaa. Käykää Bobin luona. Haluaisin nähdä sinut pian! Sinun, Matti. → *Good Bob! Thank you for the gift. That was great! You still remember we haven't met long. We're celebrating our family. Parents, brothers, sisters were there. Family things. We had one thing I'll never forget. We'll take a picture where Mom and Dad were, both my sisters and my brother together. All people in the same place. That's great. Go to Bob's. I 'd like to see you soon! Yours, Matti.*

Table 2: Illustration of the baseline pivoting method for the three source texts in Table 1. The Finnish and Swedish texts have been translated to English (in small font) and back to Finnish and Swedish (in larger font). The English text has been translated to Finnish (small font) and back to English (larger font).

out-of-domain parallel data can effectively correct a multitude of different error types in text without the explicit modeling of these errors. Some further examples of corrected errors are shown in Table 5.

This is an important finding, as language is complex and hard to handle successfully in a "silo manner", fixing typos, grammar and naturalness isolated from each other in separate steps. We initially had an idea of using existing proofing tools (spell checkers) in a preprocessing phase. However, many errors are not unambiguously spelling mistakes, as they may produce valid word forms, but which are wrong in context. We also considered an "oracle" approach for comparison, in which we would fix all the typos manually before applying our automatic methods, but it turned out difficult to decide what exactly were plain typos and how far the "oracle" would stretch.

We have chosen character-based models in order for these models to be less sensitive to noisy data. Using full words or longer word fragments would introduce numerous out-of-vocabulary words, when words in source sentences contain spelling mistakes. Comparing to Google Translate (Table 6) it seems that Google is more sensitive to noise related to typos and largely leaves such errors unfixed (for instance, the "English" words *beatufull* and *awsome*).

In line with earlier work on translation of non-native text (Anastasopoulos et al., 2019), we find that augmenting clean parallel data with artificially-introduced errors can make a system more robust and improve performance. In our case we find a discrepancy between different lan-

Hello, Mary. I'm on a vacation. I've been in China in two weeks. I already went to many cities. It was a really nice journey. Meet a lot of friends. I've eaten a lot of Chinese friends with me. That's really funny. I want to go to Shanghai and buy something about Shanghai. Hopefully, I'll see you soon! Hello, Matt. → *Hei, Mary. Olen lomalla. Olen ollut Kiinassa kahden viikon päästä. Menin jo moneen kaupunkiin. Se oli mukava matka. Tavataan paljon ystäviä. Olen syönyt paljon kiinalaisia ystäviä kanssani. Todella hauskaa. Haluan mennä Shanghaihin ostamaan jotain Shanghaista. Toivottavasti nähdään pian. Hei, Matt.*
Hey, Tove! Now I've finally come to my new town Vaasa. I'm really good, but I'm tired. The movement to the residence took four hours and they were two men who helped me wear thin furniture. The residence is light and here are big windows that give daylight to the room. There's two rooms, kitchen and WC, so there's pretty big local for me. Come and tell me next month. We're gonna call. Hot greetings for all of you. Maija → *Tove! Nu har jag äntligen kommit till min nya stad Vaasa. Jag är verkligen bra, men jag är trött. Rörelsen tog fyra timmar och de var två män som hjälpte mig att bära möbler. Bostaden är ljus och här är stora fönster som ger dagsljus till rummet. Det finns två rum, kök och WC, så det finns ganska stort lokalt för mig. Kom och berätta nästa månad. Vi ringer. Varma hälsningar för er alla. Maija*
Rakas Bob! Kiitos lahjasta. Se oli hämmästyttävää! Muistat vielä, ettemme ole tavanneet pitkään aikaan. Juhlimme perheemme kanssa. Vanhemmat, veljet, siskot olivat siellä. Perheasioita... Meillä oli yksi asia, jota en unohda. Otamme kuvan, jossa äiti ja isä olivat, molemmat siskot ja veljeni yhdessä yhdessä kuvassa! Kaikki ihmiset samassa paikassa. Mahtavaa. Käy luonamme, Bob. Haluaisin nähdä sinut pian! Sinun, Matti. → *Dear Bob ! Thank you for the gift. It was amazing! You remember that we haven't met in a long time. We're celebrating with our family. Parents, brothers, sisters were there. Family things... We had one thing I won't forget. We'll take a picture where Mom and Dad were, both sisters and my brothers in one picture! All people in the same place. That's great. Come to us, Bob. I'd like to see you soon! Yours, Matti.*

Table 3: Illustration of the baseline pivoting method augmented with artificial errors for the three source texts in Table 1. The Finnish and Swedish texts have been translated to English (in small font) and back to Finnish and Swedish (in larger font). The English text has been translated to Finnish (small font) and back to English (larger font).

	Model	1	2	3	4
en	Baseline	28.7	23.0	18.9	29.4
	+Errors	23.5	25.2	23.5	27.8
fi	Baseline	33.1	24.1	16.0	26.8
	+Errors	30.4	25.1	17.3	27.2
sv	Baseline	19.1	25.6	17.1	38.2
	+Errors	11.8	25.3	16.8	46.2

Table 4: Manual annotation results for the Baseline model and the Baseline model with error-augmented bilingual data (+Errors). Proportion [%] of generated sentences annotated as one of the four categories.

guages, with English improving the least and Swedish the most. Differences in translation quality between the different language pairs can be a contributing factor, as there is the least amount of Swedish data for training, which may explain the proportionally larger benefit of additional training data. However, we also note that the degree of inter-annotator agreement is rather low, which may call for clearer annotation guidelines in the future.

A potential bottleneck in our approach is the pivoting method. When two different translation steps are included in the pipeline, the risk of compounding errors increases, and the model is dependent on the quality of two translation directions. We leave for future work further investi-

Minä löydäin pientä koiraa kadulla. → *Löysin pienen koiran kadulta.* *Siella on hyvää kaverit.* → *Siellä on hyviä tyyppejä.* *Minä asun toisella puolella kaupungilla.* → *Asun kaupungin toisella puolella.* *1. opettaja oli kärsiväinen* → *Ensimmäinen opettaja oli kärsivällinen.*
Det finns inte länge bort här. → *Det är inte långt härifrån.* *darfor att jag glommde i dag är min mamas fodelsedag, och jag behöver går där.* → *Jag glömde att idag är min mammas födelsedag och jag måste gå dit.* *Jag hopas du är inte arg!* → *Jag hoppas att du inte är arg!* *Jag måste hjälpa min systorna på en school-projekten.* → *Jag måste hjälpa mina systrar på ett skolprojekt.* (Here a small error remains. The preposition should be: ... ***med*** *ett skolprojekt.*)
Other chance is not. → *There's no other chance.* *Depence if you take with new tires or with old one's.* → *Depends if you take new tires or old tires.* *They have very funny.* → *They're having fun.* *The first you must to talk to these young people for this problem.* → *First you have to talk to these young people about this problem.*

Table 5: Examples of successful corrections of sentences in the test data. These translations have been produced using the baseline pivoting approach. Typical errors in the Finnish input are incorrect inflections, incorrect word choices and omissions of umlauts. Similar errors occur in the Swedish data with additional challenges related to word order, agreement and foreign words. The native language of the authors of the English sentences is revealed by the Finnish sentence structure of the English sentences.

gations into how monolingual data could be used effectively to circumvent the need for pivoting and increase performance.

In addition to fixing obvious grammatical errors in source sentences, we find cases where our model introduces fluency edits. This can be seen, for instance, in more idiomatic choices of words: it is typical for non-native Swedish speakers to use the verb *finna*, which resembles *to find* in English and means the same thing, but a more natural choice would be the verb *hitta*. Our models do change *finna* to *hitta*. Similarly, the Finnish expression *mennä takaisin* (go back) is replaced by *palata* (return).

Together our results show promise for using a standard NMT approach to improving and paraphrasing noisy language learner text. As test data we have used Finnish, Swedish and English portions of the YKI corpus, which to our knowledge have not been studied in this setting before, and could be of special interest to a Nordic audience. As far as computer-assisted language learning is concerned, we find the fluency edits introduced by the models especially encouraging. The models go beyond simple grammatical error-correction and can help language learners improve their skills toward more fluent and native-like language production. We believe that our approach is particularly beneficial to more advanced learners, who want to be able to use their new language more autonomously, in situations where no human teacher is available.

Acknowledgment

The authors wish to acknowledge CSC, the Finnish IT Center for Science, for providing the computational resources needed to carry out the experiments in this study.

References

Antonios Anastasopoulos, Alison Lui, Toan Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. In *Proceedings of NAACL-HLT 2019*, pages 3070–3080, Minneapolis, Minnesota.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Fatemeh Behjat. 2011. Teacher correction or word processors: Which is a better option for the improvement of EFL students writing skill? *Journal of Language Teaching and Research*, 2(6):1430–1434.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Moi Mai: Olen Kiinan loma-matka. Olen ollut Kiinassa kaksi viikoonia. Menin jo moniin kaupunkeihin. Se oli todella mukava matka. Nähdään paljon ystäviä. Olen syönyt paljon kiinalaisia ystäviä herkuisetiin. Todella hauskaa! Haluan edelleen käydä Shanghaissa ja ostaa jotain Shanghaissa. Toivottavasti nähdään pian! Ystävällisin terveisin Matti
Hejsan Tove! Nu har jag äntligen kommit till min nya stad Vaasa. Jag känner mig väldigt bra men lite trött är jag. Flyttet till hemmet tog fyra timmar och de var två män som hjälpte mig att bära tunna möbler. Boendet är ljust och här finns stora fönster som ger dagsljus till rummet. Det finns två rum, kök och toalett, 56 m så ganska stort rum för mig. Kom och se mig nästa månad. Vi ringer. Varma hälsningar till er alla Maija
Good Bob! Thank you for the gift. It was beatufull! You still remember that we haven't met for a long time. We celebrate with our family with our family. Parents, brothers, sisters were there. Family matters ... We had one thing I would never forget. We take a photo of mother and father, both sisters and brothers together in one picture! All nations in the same place. Awsome. Visit us Bob. I'd like to see you soon! You, Matti

Table 6: Applying the pivoting method using Google Translate (as of June, 2019). The Finnish and Swedish texts have been translated to English and then back. The English text has been translated to Finnish and back. In comparison to our own results in Tables 2 and 3 it is not obvious which method is the most effective, as Google Translate does not seem to cope well with errors in the source sentences.

Mathias Creutz. 2018. Open Subtitles Paraphrase Corpus for Six Languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montreal, Canada.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Mariano Felice and Ted Briscoe. 2015. Towards a standard evaluation method for grammatical error detection and correction. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 578–587, Denver, Colorado.

Simon Flachs, Ophélie Lacroix, and Anders Søgaard. 2019. Noisy channel for low resource grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 191–196, Florence, Italy. Association for Computational Linguistics.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2018. Near human-level performance in grammatical error correction with hybrid machine translation. In *Proceedings of NAACL-HLT 2018*, pages 284–290, New Orleans, Louisiana.

Mika Hämäläinen. 2019. UralicNLP: An NLP library for Uralic languages. *Journal of Open Source Software*, 4(37):1345.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of NAACL-HLT 2018*, page 595606, New Orleans, Louisiana.

Satoru Katsumata and Mamoru Komachi. 2019. (Almost) unsupervised grammatical error correction using synthetic comparable corpus. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 134–138, Florence, Italy. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Krister Lindén, Erik Axelson, Senka Drobac, Sam Hardwick, Juha Kuokkala, Jyrki Niemi, Tommi A Pirinen, and Miikka Silfverberg. 2013. HFST – a system for creating NLP tools. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 53–71. Springer.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 588–593, Beijing, China.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Gustavo Henrique Paetzold and Lucia Specia. 2016. Collecting and exploring everyday language for predicting psycholinguistic properties of words. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 669–1679, Osaka, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania.

Eetu Sjöblom, Mathias Creutz, and Mikko Aulamo. 2018. Paraphrase detection on noisy subtitles in six languages. In *Proceedings of W-NUT at EMNLP*, Brussels, Belgium.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.